

# Characterizing Structure in Cross Site Scanning Dynamics

*Scott Campbell, Craig Lant*

*National Energy Research Scientific Computing Center*

[scottc@nersc.gov](mailto:scottc@nersc.gov), [clant@nersc.gov](mailto:clant@nersc.gov)

## Abstract

In this work we show the existence of inter-site structure in network scanning behavior from the perspective of a number of laboratory and university networks. Once such structure is demonstrated, we measure a number of its characteristics. Using connection records from each site and processing them independent of any local infrastructure, we are able to track the movement of individual IP addresses with high time resolution across the monitored address space. This allows us to draw on an unusually clean set of scan related information both in terms of algorithmic consistency and the degree of detail that we can extract from individual scanning events.

## 1 Introduction and Motivation

The passive analysis of network traffic – both actively hostile and radiative – can be grouped together into three general types: radiation, telescope and scanning analysis. Classic passive traffic analysis has generally focused on the aggregate behavior of sections of address space and the traffic that is destined for them. Radiation style analysis focuses more on the average behavior of traffic as seen by the observing network(s), creating general descriptions of connection size, duration and other salient characteristics [6] [7] [8]. Individual addresses and timing data are typically ignored or reduced to a statistical description. Network telescope measurements use large swaths of address space to derive behavior descriptions of distant addresses - the larger the collection of address blocks, the greater the aperture of the telescope. While individual addresses might be observed, the packets are not characteristically directed at a given host and that any non-local timing information is lost by the geometric distribution built into the traffic model[14]. Scanning analysis tends to be focused on an individual site, and statistical generalizations are used to describe aggregate behavior rather than individual address behavior [15].

Understanding large scale behavior of hostile scanning across multiple networks has value both in terms of research as well as practical applications. Characterizations of network scanning activity is widely used for everything from firewall design to intrusion detection. A better understanding of cross-site activity can be used to improve the effectiveness of devices, processes, and algorithms based on these measurements. Furthermore, knowing how scanners tend to group their attacks as well the direction and velocity of those scans can be used immediately to improve inter-site coordination of incident response by introducing conditions on who should communicate and how fast the interactions need to be.

## 2 Approach

In this study, methods of identifying clusters of destination network targeted by individual scanners are presented and tested against one another. Once such clusters (or structures) are identified, their characteristics are measured and compared against non-structure address spaces. The problem of multi site scanning takes place in two steps. The first looks at the distribution of hostile scanning addresses within a set of destination networks to see if there are stable groupings which might be used to infer inter-network and inter-site ‘structure’. For the remainder of the paper, the term ‘structure’ will refer to this definition. Once the existence of this structure is shown, various characteristics are measured and compared to the non-structure networks.

To address the first part of the design, we developed a tool that would be able to digest text based connection style data in a site agnostic way and identify external scanning addresses based on an arbitrary algorithm. Since not all of the networks providing data were willing to do so in a non-anonymized way, it was also necessary to address local privacy concerns. To take care of these concerns, a filter was implemented to avoid recording local site addresses. In addition, an option was installed which would record only the MD5 hash of any reported addresses. This was not designed to deter a concerted attack on the obfuscated addresses, but rather to prevent trivially identifying the attacking addresses.

To ensure portability, data is stored in line delimited text files which can be processed by off the shelf tools such as shell scripts and scripting languages. Analysis is done in as automated and repeatable manner as possible based on suggestions by Paxton [13]. More details on this will be found in section 3.

This choice for scan detection algorithms is based Jung's [2] analysis, which demonstrated low false positive rate relative to a number of other algorithms for the native Bro as well as the Threshold Random Walk (TRW) algorithms [2]. For the native scan detection algorithm, no changes were made to the design. For TRW, a small change was made to report source addresses which are classified as non-hostile as well as the more typical hostile logging.

### 2.1 Initial Structure Detection

To determine the existence of temporally stable structure, a series of tests were run measuring the overlap of several quantities found in the scan detection measurement. Each data point in the scanning set can be guaranteed to contain at least the following values: initial timestamp, source address and destination network. For each of the sample measurements one of the three values will be used as the dependent variable. Three independent tests are run against the data since the exact nature of the structure is unclear.

Since participating sites can contain multiple, non-contiguous networks, a *network* will be the unit of measure in this study rather than a site. A network is a routed block of address

space publicly associated with a given site via the usual Internet BGP mechanisms. When referring to the physical location where a set of local networks are homed, the term *site* will be used.

### 2.1.1 Network Crossover

In this test, the overlap of scanners common to pairs of networks is used as a measure for describing structure subcomponents. To calculate the address overlap between a pair of networks, scanners are identified for each network over a given time interval. The magnitude of the intersection between the two scanning sets is divided by the smaller of the two values. This is described by equation 1:

$$O_{AB} = \frac{|A \cap B|}{\text{MIN}(|A|, |B|)} |_{\Delta t}$$

**Equation 1**

The rational for selecting the minimal value is to give weight to smaller net blocks since there is no guarantee that blocks  $A$  and  $B$  will be the same size, and the number of scanners per unit time is proportional to the size of the network. In addition, limits need to be set for the minimum size allowed by this test. If too small, any changes in the number of overlap addresses will have an inordinately large sway in the final overlap value. For our tests, the value of  $C(S_n, S_m)$  is averaged over each of the time intervals with the standard deviation providing some sort of confidence measure.

To use this method to measure structure, pair wise sets are calculated in table form to express the average overlap and standard deviation of each data pair. For example, given a set of sites to look at –  $\{S_1, S_2, S_3, \dots, S_n\}$ , the crossover between identified hostile can be expressed in a classic ‘mileage chart’ format:

$S_1$	-----			
$S_2$	$C(S_2, S_1)$	-----		
$S_3$	$C(S_3, S_1)$	$C(S_2, S_3)$	-----	
$S_n$	$C(S_n, S_1)$	$C(S_n, S_2)$	$C(S_n, S_3)$	-----
	$S_1$	$S_2$	$S_3$	$S_n$

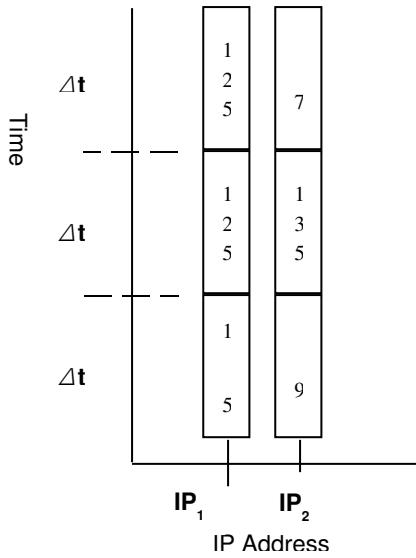
**Table 1**

Since the test is symmetric – i.e.  $C(S_3, S_2) = C(S_2, S_3)$ , only half of the table need to be calculated.

### 2.1.2 Temporal Locality

Looking at the set of scanning data using a sliding time box approach provides another way of clustering the data based on temporal locality rather than destination network overlap. By mapping scanner source address and initial timestamp pairs to the destination network, it is possible to infer that an IP address scanned a combination of destination networks within a given time window. See figure 1 for an illustration. In the case described by figure 1, during the time

window  $\Delta t_3$  IP address 1 was identified as a scanner for networks 1, 2 and 5. By taking a tallied summary of networks seen for the entire time and network range, it is possible to create a list of networks more likely to be seen together.



**Figure 1**

addresses this, and suggests a 24 bit masking to identify locality meaningful addresses.

Venkataraman [11] describes a similar relationship as a spam reduction technique – in that case the notion of local is described by the AS size which would be difficult to fold into this technique. A representation of the technique would be quite similar to figure 1, but the IP addresses would be replaced with an address range,  $\Delta IP$ .

### 2.1.3 Address Crossover

By ignoring the time boundary in the crossover calculations described in 2.1.1 and looking at all the data at the same time, another perspective can be arrived at. This test is conceptually simpler than either of the previous two, but lacks any significant resolution in terms of network and time dynamics. Like 2.1.1, the proportion of scanner overlap will be expressed in the same ‘mileage chart’ format.

It is worth noting that we expect to see different results in this measurement than those arrived in for 2.1.1 since the overlap window has been completely removed. The effect of this is to magnify the degree of overlap since networks scanned weeks apart by the same address will be counted the same as if the scanning happened minutes apart.

## 2.2 Quantitative Description of Structure

Provided that some sort of structure exists within the destination networks, it seems natural to try to quantitatively measure some of its characteristics. In this section two general qualities will be measured. The first involves a simple mechanical interpretation of the data.

For this exercise, it is interesting to evaluate the clustering with a variety of  $\Delta t$  values in order to test locality. If the proportion of pairings does not change much with increasing  $\Delta t$  values then networks are being scanned closely together. This would be a very useful number for determining how fast information needs to be shared between sites that typically experience common scanning sources.

In addition to changing the values for  $\Delta t$ , there has been some work in predicting the behavior of an address based on the behavior of addresses located near it. Collins et. al. [10] directly

Directional bias and inter/intra scanning velocity will be looked at for addresses scanning the identified structure as well as a control group. The second measured quantity looks at differences between random radiation scanning and directed scanning as a function of time for both structure and non-structure address sets. Details for each can be found in section 2.2.1 and 2.2.2 .

In order to provide a set of linked destination networks for the structure definition, the top five identified crossover candidates will be used. These networks are chosen from the individual pairs are identified by the pair-wise test described in 2.1.1, and the multi-network structures are clearly described by the time crossover method. By looking at the relationship between these networks, it is interesting to note that the majority of nodes share partners with one another.

### **2.2.1 Directional Bias and Velocity of Scanners**

In order to infer a bias in the direction of scanning, we look at the average behavior of addresses as they cross the various network ranges. When an address is identified as a scanner, a timestamp is recorded into the logs. A vector is created holding the time values for each network where the address was seen with each site as a dimension. Dimensions are then sorted by the numerical ordering of the initial IP address in the network range. For example, looking at four sample networks the vector would look like  $\mathbf{v} = \langle t_1, t_2, t_3, t_4 \rangle$ . To identify directional bias, the vector components are traversed - if  $t_1 < t_2$ , there is a *right* bias in the direction. If not, the direction is considered *left* biased. This process is continued along the length of the data vector.

In order to test the drift bias, a control group needs to be created. This control group is composed of addresses identified as scanners which have crossed multiple networks, but are not in the set of addresses which have scanned networks within the structure. In order to better understand the drift bias as a function of the number of destination networks visited, the control group will also be broken down by network count as well.

The idea of network velocity, uses hosts scanned per second as a basic metric. This quantity can be thought of both in terms intra-network and inter-network rates. This measurement augments the directional bias information already calculated, and to gain greater insight into differentiating attacks directed at the destination network set from non-directed background radiation type attacks.

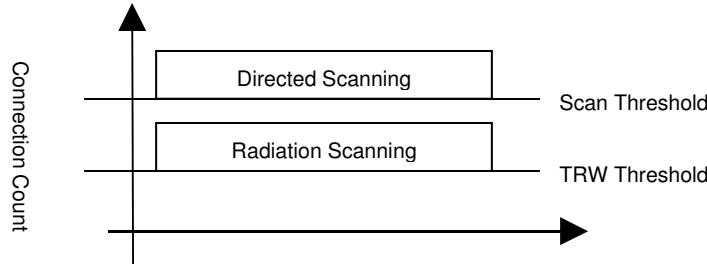
Using the same set of addresses as the direction bias check, a vector is created which holds the initial and final contact time stamp and the number of unique local addresses contacted by the hostile address for each of the ordered destination networks. If an attacker visits a network multiple times, only the initial contact data is used. For the intra network velocity, the rate of scanning is approximated by taking the total number of addresses and dividing it by the duration of the scanning.

For inter network measurements, the problem is a little different. Since this is an approximation of what the rate of scanning must be for a linear traversal of the address space between two networks, the numerical difference between the first address in each network block is used for the number of addresses and the differences between first contact times is used as a duration.

More details and data analysis can be found in section 4.2.2 .

### 2.2.2 Measuring Directed vs. non-Directed Attacks

The next characteristic looked at measuring the ratio of radiation scanners vs. directed scanners for hosts inside and outside the structure address set. Differentiating between the two scanning types necessitates an arbitrary distinction. We assume that radiation like scanning is caused by attacking systems randomly connecting to destination addresses. This can be thought of as a series of Bernoulli trials expressed in a classic geometric distribution. [14] Directed scanning takes an address range and connects only within that range. An example of random scanning would be the Slammer Worm [5], while a nessus [4] scan would be an example of a directed scan.



**Figure 2**

In order to differentiate between the two groups, it is necessary to understand the distinction between the native and TRW scan detection algorithms. An assumption here is that randomly driven scanning will create a smaller number of connections per unit time than directed scanning for a given network. If an address is identified as a scanner by TRW and not by the native algorithm for a sufficiently long time duration, then the scanner source is considered the product of radiative scanning. If the native algorithm identifies it as well, then it is considered a directed scan. There are a handful of cases where the native algorithm detected a scan and TRW did not. This can be readily understood by noting that a sufficiently large number of successful connections were made to force a 'benign' solution for the TRW test. This set of addresses are treated as a regular directed scanners.

## 3 Design and Implementation

In order to generate a data set which reflects a consistent image of scanning activity across multiple networks, it is necessary to simulate scan detection by the direct analysis of connection records. This sidesteps issues with multiple threshold values being used for scanning

definitions across different sites as well as providing a way to separate local, non-contagious networks into discrete units where they are often monitored together.

In order to achieve this, a series of template scripts were written which could be trivially modified for a site's local network. These scripts are designed to be simple and transparent, making it easy for individual sites to verify what they are doing and alleviate any concerns they might have about running them. Since the scripts are agnostic with respect to the format of the data they're given, it should be possible to use connection records for different sources without too much difficulty as long as the same general notion of bidirectional connection records is maintained.

Script output includes line delimited records containing both native and TRW based scan detection. Record types and their details are as follows:

Record Type	Record Details
Native Bro Scan	initial packet time, time to trigger, time to complete, number of scanning hosts, number of source ports
TRW Scan	initial packet time, time to trigger, destination port
TRW Benign	initial packet time, time to trigger, destination port

Table 2

The “TRW Benign” implementation represents the second solution allowed by the TRW test. This solution identifies when a connection is deemed non-hostile – for example when a host makes a large number of successful connections. In Bro these values are not normally recorded, but the additional data could provide some insight into behavior which is driven by the application layer.

To ensure that the results of this research are accurate and easily reproducible, the processing and organizing of the data received from each site was also automated. This makes it very easy to add new data from additional sites or to simply move or resize the time window used.

## 4 Evaluation

Since the raw data records are stored as line delimited records, scripts can be used for simple parsing and analysis. These tools can provide a way to interactively explore various analysis scenarios. This is further helped by the use of automated scripts for loading, cleaning and processing the raw data files delivered by the various sites.

In formatting the raw data files, several conditions are introduced. First a time unit of 24 hours is used to divide the data into more usable chunks. Since there is no reason to assume that a 24 hour log rotation cycle is used by all sites, all scan detection records for a given network were sorted by time and broken up into blocks of 24 hours starting at midnight. This may introduce systematic errors if some sites reset state on 24 hour cycles while others do not.

Second, all timing data was zeroed to PST time in order to allow for simple comparisons independent of which site is being looked at.

We also made the assumption that timestamp data was accurate to within bounds provided by typical Linux or BSD NTP implementations. NTP drift time was measured between two hosts for a 12 hour period, but the effect was nominal relative to the size of the time values being used in this analysis. Time drift was looked at by Murdoch in [16], and results also enforce the notion that system timestamp drift is negligible relative to the values being looked at.

The final network distribution was 9 class B networks, and 12 class C networks spread over 5 different sites made up of DOE Labs, NSF computational facilities. In a later version of this study we are hoping to add more educational, commerce and non-US address ranges to provide a more complete view of network scanning as there is evidence to suggest that US governmental network ranges are being excluded from many attack lists. We recognize that the distribution of networks used in this survey may skew the structural identification and will address this in later versions.

## 4.1 Initial Structure Detection

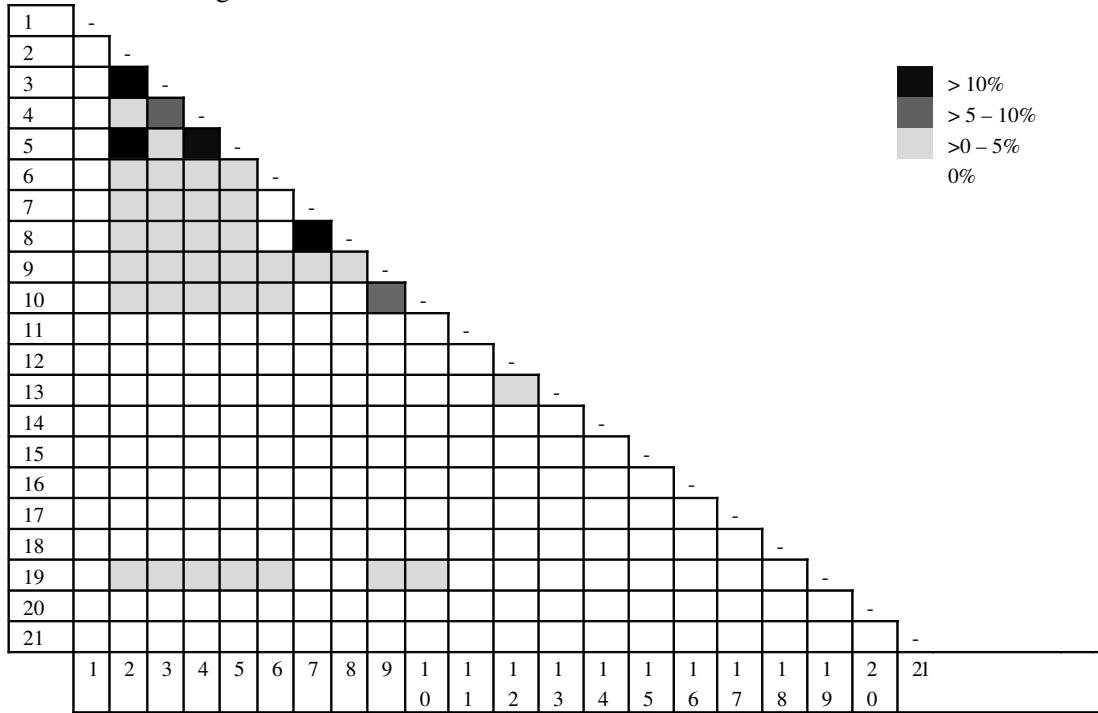
For several of the evaluation techniques, the hope was to use a standard deviation as a way to measure the quality of the individual data points in pair wise sets. What we found was that the data distribution is non-Gaussian, so classic standard deviation calculations do not return meaningful information. An example of this would be network crossover data, described in section 2.1.1. These values can vary significantly over time and in many cases are double peaked. This is likely due to a continuing variation of the attack traffic.

### 4.1.1 Network Crossover

In section 2.1.1 a test was presented describing the scanner source overlap between two destination networks. When this test is applied to the total collection of network pairs, the results can be found in figure 3 below. Initial results show that the overlap distribution is non-consistent across different destination networks which is an indicator of possible structure within the pairs of destination networks.

The data points in figure 3 represent the crossover between the two networks indicated in the axis values. Final values are arrived at by taking the average of the individual crossover values (which are calculated from 24 hour periods over the entire sampling period). A cross-plot of network pairs shows significant variation in the percentage of traffic seen by each set.

Average Percent Crossover for Network Pairs for 24 Hour Window



**Figure 3**

In order to reduce unreasonable thrashing in crossover values, a minimum threshold of 5 data points was required for the denominator in equation 1. Less than that and the small value would dominate the equation.

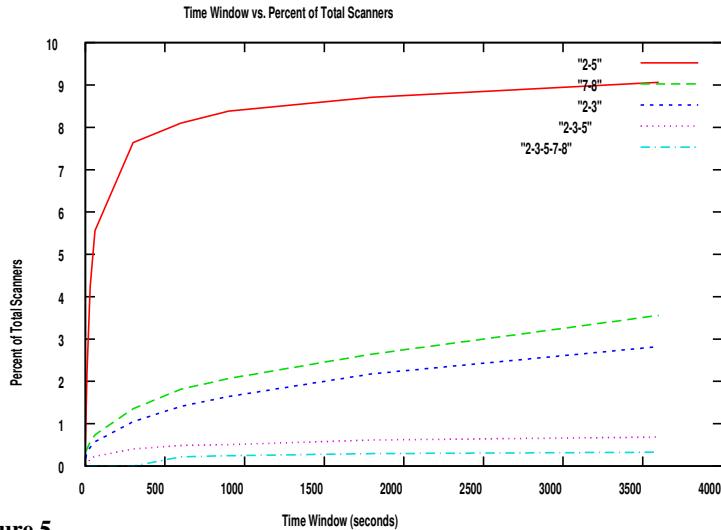
#### 4.1.2 Time Crossover

The second comparison test tracks scanning source addresses based on how many destination networks they are identified scanning during a time window. The distinction between this and the destination network mapping in 4.1.1 and 4.1.3 is principally about what is being measured. In the destination network tests, this is the proportion of shared scanning addresses between two networks. In the time window test there is no limitation on the number of networks evaluated – if an address is identified as a scanner by one or several networks within the time window, so any network combination pairs in the time window are assigned as a possible structure. By aggregating the values across the set of time slices, a set of destination networks can be identified.

A sorted table of network combinations (ie ignoring windows where only a single network was scanned), shows a exponential distribution of combination counts. Combination count here means the a list of network combinations sorted by the number of times they appear. The final five values represent less than 3% of the total possible combinations, yet account for greater than 93% of the total count. This distribution is represented in figure 4.

**Figure 4**

Looking at figure 5, we see a representation of the top five destination pairs across the set of time window. The interesting thing to note with this is the data distribution along the x-axis as the size of the time window increases. This graph is similar to results by Sachin et. al. [9] where the vast majority of interactions seem to fall within a short ( $< 750$  sec) time window. The most significant difference is that the time window is longer than that provided by Sachin. In addition it is worth noting that the collective proportion of the identified 'structure' scanners is less than 18% of the total.



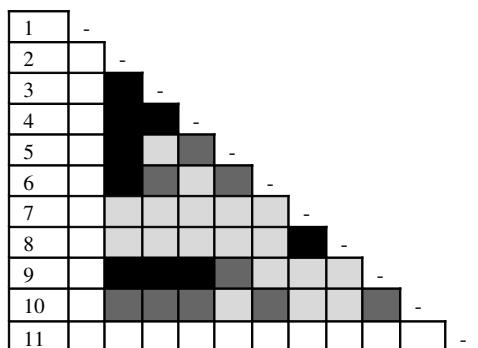
**Figure 5**

causal linking between the network entries in each of the address x time boxes.

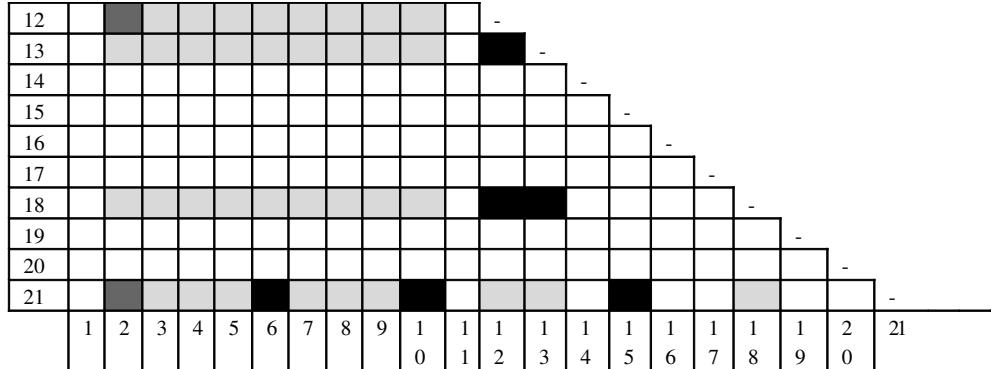
#### 4.1.3 Total Address Crossover

In this test, the same technique was used as 4.1.1, but the individual 24 hour time segments are ignored. This test is done to see the maximum possible crossover between network pairs – temporal fluctuations like new worms and network outages tend to be smoothed over and the larger overlap characteristics stand out. Unfortunately with the larger data set there tends to be a little more obscuring of the fine features that would stand out in the higher time resolution analysis in 4.1.1.

##### Average Percent Crossover for Unlimited Window



As previously mentioned, this method of structure detection provides for the direct discovery of more complex groupings than simple pairings, albeit at the cost of some ambiguity. For example, the sets {2,3}, {2,5} and {2,3,5} are all independently described in figure 3. While there is some expectation that statistical results limit artificial groupings, there is no strong



**Figure 6**

The significant differences between figure 6 and 3 include added groupings of high crossover caused by the longer sampling period.

This slice of the data set also provides a nice opportunity to see the distribution of the total number of networks touched by individual scanners over the entire sample period. Looking at table2, the distribution of network count over sample period is provided.

Number of Unique Networks	Count	Number of Unique Networks	Count
1	146716	7	122
2	73152	8	109
3	8679	9	29
4	4351	10	5
5	2381	11	3
6	642		

**Table 2**

## 4.2 Quantitative Tests of Structure Properties

Looking at the network sets described by each of the testing methods – particularly network and time crossover – it is possible to come up with a subset of observed networks which are agreed upon by all the tests. In this case it is networks {2,3}, {2,5}, {7,8}, {2,3,5} and {2,3,5,7,8}. These networks are chosen since the individual pairs are identified by the pair-wise test described in 2.1.1, and the multi-network structures are clearly described by the time crossover method. Given the limitations imposed by the total address crossover particularly for time granularity, the results from it were not used in the definition.

Assuming that the description of the structure is accurate, a series of measurements are made contrasting characteristics between the network traffic as seen within the structure with that outside. The first two are a mechanistic interpretation of the data – bias in scanning direction and velocity distributions. The last looks at the ratio of directed scanning vs. non-directed radiation type scanning.

### 4.2.1 Direction Bias and Scanning Velocity

The first thing to look at is directional bias in what networks are selected from within the scanning structure. To do this, the timestamp from initial contact for each of the identified scanners is recorded. When placed into a data structure sorted by IP address, it is possible to infer directionality of scanning by comparing the time values for position  $i$  and  $i + 1$ .

Starting with the entire set of addresses identified as scanners for the address ranges identified in section 4.2, the above method is used to calculate bias for approximately 387,000 transitions between networks.

Looking at direction from structure group (table 3), we see a clear bias from lower numbered addresses to higher. The data in the second row consists of an absolute count and the proportion represented as a percent. While this is consistent with intuition, it also enforces the notion that scanning is a non-random phenomenon.

Left	Right
179858 (.464)	207263 (.536)

**Table 4**

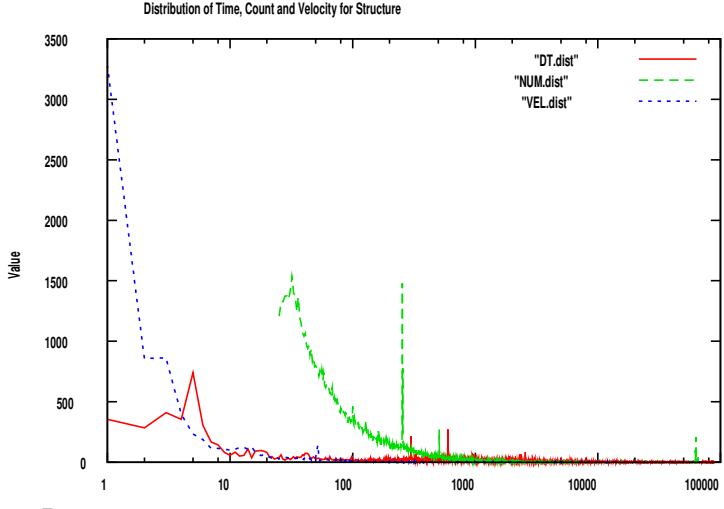
As described in 2.2.1, the set of control addresses are those scanners who crossed more than one network, but which are not found in the set of addresses from the structure group. This intersection operation may skew the control group behavior by removing scanners who might otherwise be counted for random behavior or structural actions which escaped the initial analysis here. Table 5 provides bias information for the number of networks identifying individual attackers found in the control group.

Minimum Instances	Left	Right
6	118 (.480)	128 (.520)
5	194 (.430)	257 (.570)
4	357 (.430)	473 (.570)
3	839 (.626)	502 (.374)

**Table 5**

In both cases it is clear that the distribution is a significant deviation from what would be expected if there were no bias in the directional choice.

The velocity analysis is a measure of the rate of IP address scanning for local intra-network as well as large scale cross (intra) network behavior. For the intranet values, the number of unique hosts contacted, the time duration and the ratio of the two values is presented in figure 7. The data for these measurements is directly provided by the net scanning record. The number of hosts and the duration are provided as separate values since they can individually provide a considerable amount of information about the dynamics. The spike for the number of hosts scanned (NUM.dist) is at 254 which is not particularly surprising.



**Figure 7**

A significant interest here is the relatively low number of addresses that are typically contacted per unit time, as well as the short window that the destination networks are exposed to. These values are consistent with the host fanout values described in Allman et. al. [15].

Measuring velocity values for inter-network traffic requires a few changes. While the address range in intra network scanning is well defined and compact, the inter-network distances tend to be two or more orders of magnitude larger. There is also a level of uncertainty in the time values. The same IP address might be scanning a set of networks at various times so it is not always clear when two

measurements are related. For purposes of this paper, we used the closest two time values that would make a set. For example, if an address was seen scanning networks A and B at times  $T_{A1}$ ,  $T_{A2}$ ,  $T_{B1}$ , the time difference would be  $|T_{B1}-T_{A2}|$  since that provides the closest fit.

In order to get some notion of the relative velocities seen in inter and intra network scanning, we looked at the behavior over a single pair of class B networks. In this case, NET2 and NET3. We recognize this as being potentially non-representative of velocity distributions in general, but the results are still interesting. Calculations for velocity are done the same as before. For intranet velocity, the number of addresses scanned is divided by the duration of the scan. For internet calculations, the IP distance between the two subnets is used for the number of hosts and the difference in first contact time is used for the time value.

Results for the internet values follow the same general slope as intranet values, but then begin to climb back up as the velocity increases. This is readily explainable by noting that the  $\Delta t$

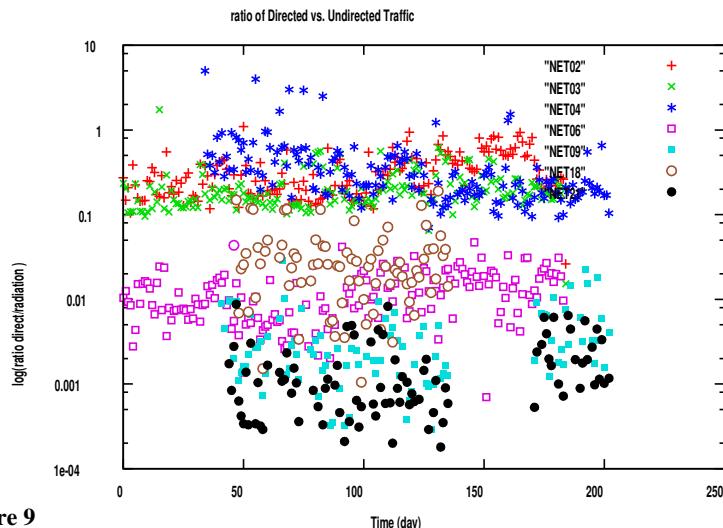
value for these high velocities is quite small. Such small time values would be explained by the structured scanning discussed in the first part of this paper.

Looking at figure 8, we see a spike in the internet velocity. This seems to be an artifact in the analysis process caused by the 24 hour checkpoint cycle. If the distance between the networks is divided by 24 hours, you get a value exactly equal to that of the spike. This is most likely caused by addresses which are continuously scanning, or whose cycle time is greater than 24 hours for both networks at the same time. Compared to the remainder of the graph, the proportion under that part of the curve is quite small.

**Figure 8**

#### 4.2.3 Directed vs. Radiation Scanning

The final measure is the ratio of directed vs. radiative scanning. As described in 2.2.2, the TRW scan algorithm is used here to measure non-directed scanning which would otherwise fall beneath the threshold required for the native scan detection. Since data representation for multiple networks can be complex, the ratio values for a number of networks are plotted as a function of time. This will show both the relative stability of the metric over time as well as allowing groupings to form without bias.



**Figure 9**

In this case, there is a reasonably clear separation between the set of networks which have been included in the structure group and those who have not. A higher value in this case indicates that there is proportionally more directed scanning. Here NET02, NET03 and NET05 are representative of the structure we are looking at.

There are several things that should be noted in the graph. First NET7 and NET8 are not present since they did not collect TRW information, hence it was not possible to run this test on those networks. The second is the possible role of the host density on the destination network side and its roll in possible biasing the TRW results. If a network is densely populated, you would expect a slight increase in the rates of failures for TRW scan detection. For the networks looked at, this is not expected to be a problem.

An interesting observation can be made about NET18 in figure 9. That value seems to be the single instance which violates the separation between the two groups. Looking at figure 6,

you can note that this net had significant crossover with several other networks, but with a time window which exceeded the one used for structure testing and verification.

## 5.0 Related Work

The analysis of hostile traffic has historically been a rich field of research. In this section we will survey a cross section of related work, both in terms of algorithms and tools used to implement them. In addition, work relating to structural analysis will be addressed as well.

In terms of general scanning analysis, the native Bro [1] scan algorithm (BSA), a host is allowed to have a set number of *failed* connection attempts before they are identified as a scanner. The Threshold Random Walk (TRW) [2] algorithm works by using an 'oracle' to determine if a connection will succeed or fail. A successfully completed connection drives a random walk upwards, a failure to connect drives it downwards. By modeling the benign traffic as having a different (higher) probability of success than attack traffic, TRW can then make a decision regarding the likelihood that a particular series of connection attempts from a given host reflect benign or attack activity, based on how far the random walk deviates above or below the origin. [3]

A principal tool for the investigation of traffic destined to routeable, but typically unused portions of address space is called a Network Telescope. Examples of traffic include undirected scanning and worm traffic. These tend to be driven by linear address traversal, or random/pseudo-random target address schema's [4] [5]. As described in [6], observation of traffic is based on looking at address or packet distribution – when an IP address is chosen at random, the probability of observation is expressed as a geometric distribution, while the odds of multiple packet observation are binomial in nature.

Traffic characterization is broken up into several classes, Real (non-radiation/worm) traffic modeling was classically described by Paxson in [7]. Of significance is that type of traffic described was completely legitimate (in a TCP sense) and directed. Recently the idea that connection arrivals for interactive, user driven activity are the only successful candidates for Poisson description was looked at by Karagiannis et al. [8] in terms of appropriate time scales, aggregate traffic flows and packet inter arrival time distributions.

Structural analysis of hostile activity is a way of generalizing source or destination traffic behavior. This can be in terms of patterns in attacker source address, or clusters of destination networks. Related to attack structure, Sachin's [9] excellent paper describes a similar pattern of attack structure using a combination of pure scanning and signature detection. Our study can be differentiated from it both in terms of the single connection oriented data source as well as the focus on describing characteristics of the scanning structure.

Determining intent based on previous hostile activity near a source address has been looked at in terms of botnet activity by Collins [10] and spam source Venkataraman [11]. In each of these cases, some additional benefit was gained by this modification.

## 6.0 Conclusion

In this paper we show the existence of inter-site structure in network scanning behavior using several tests. Once such structure is demonstrated, a number of its characteristics are measured including drift bias, velocity and the proportion of directed vs. non-directed scanning events. From this we see that there is a statistically significant bias in scanning direction. In addition, measured scanning velocities are consistent with other research. Finally the ratio of directed vs. non-directed scanning is measurably different between networks located within the structure compared to those not.

## 7.0 References

- [1] V. Paxson, Bro: A System for Detecting Network Intruders in Real-Time, Computer Networks, 31(23-24), pp. 2435-2463, 14 Dec. 1999.
- [2] J. Jung, V. Paxson, A. Berger, and H. Balakrishnan, Fast Portscan Detection Using Sequential Hypothesis Testing, Proc. IEEE Symposium on Security and Privacy, May 2004.
- [3] Nicholas Weaver, Stuart Staniford, Vern Paxson. Very Fast Containment of Scanning Worms, Proc. USENIX Security Symposium, August 2004
- [4] Renaud Deraison, Ron Gula, “Using Nessus to Detect Wireless Access Points”, Tenable Network Security, May 5, 2003, <http://www.tenablesecurity.com/images/pdfs/wap-idnessus.pdf>.
- [5] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford and N. Weaver, Inside the Slammer Worm, Security and Privacy, July/August 2003.
- [6] R. Pang, V. Yegneswaran, P. Barford, V. Paxson and L. Peterson, Characteristics of Internet Background Radiation, Proc. ACM IMC, October 2004.
- [7] V. Paxson and S. Floyd, Wide-Area Traffic: The Failure of Poisson Modeling. IEEE/ACM Transactions on Networking, Vol. 3 No. 3, pp. 226-244, June 1995.
- [8] T. Karagiannis, M. Molle, M. Faloutsos, A. Broido, A Nonstationary Poisson View of Internet Traffic, INFOCOM, Hong Kong, 2004.
- [9] Collaborating Against Common Enemies, SachinCollaborating Against Common Enemies, SachinKatti, Balachander Krishnamurthy, Dina Katabi, Proc. of ACM SIGCOMM IMC 2005.
- [10]M. Patrick Collins, CERT Network Situational Awareness Group; Timothy J. Shimeall, CERT Network Situational Awareness Group; Sidney Faber, CERT Network Situational Awareness Group; Jeff Janies, CERT Network Situational Awareness Group; Rhiannon Weaver, CERT Network Situational Awareness Group; Markus De Shon, CERT Network Situational Awareness Group: Using Uncleanliness to Predict Future Botnet Addresses, IMC October 2007.
- [11]Shobha Venkataraman, Subhabrata Sen, Oliver Spatscheck, Patrick Haffner, and Dawn Song. Exploiting Network Structure for Proactive Spam Mitigation, USENIX Security Symposium, Aug 2007.
- [12]D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford and N. Weaver, The Spread of the Sapphire/Slammer Worm, technical report, February 2003.

- [13]V. Paxson, [Strategies for Sound Internet Measurement](#), Proc. ACM IMC, October 2004.
- [14]David Moore, Colleen Shannon , Geoffrey M. Voelker, Stefan Savage, Network Telescopes: Technical Report,
- [15]M. Allman, V. Paxson, and J. Terrell, [A Brief History of Scanning](#) , Proc. ACM IMC, October 2007.
- [16]Steven Murdoch, Sampled Traffic Analysis by Internet-Exchange-Level Adversaries, 13th ACM Conference on Computer and Communications Security (CCS), Alexandria, Virginia, USA, 30 October–3 November 2006.